Cass Business School
City of London

Faculty of Actuarial Science and Statistics

The Neighbourhood Health Economy

A Systematic Approach to the Examination of Health and Social Risks at Neighbourhood Level

Dr Les Mayhew

Actuarial Research Paper No. 144

December 2002

Sir John Cass Business School, 106 Bunhill Row, London EC1Y 8TZ

Telephone +44 (0) 20 7040 8470
www.cass.city.ac.uk

ISBN 1 901615 66 9

"Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission".

The Neighbourhood Health Economy

By Les Mayhew

Abstract

The Neighbourhood Economy (NHE) offers a different approach for analyzing health, social needs and inequalities at local level. The techniques appear to fit very well with current Government thinking and policies in which initiatives are increasingly being targeted at smaller areas and involve multi-agency co-operation. Such initiatives require a strong evidence base but analysis using official statistical sources is heavily constrained by poor quality and coverage of data at small spatial scales and limitations of customarily used statistical methods. Because the methodology here is founded on household based data it by passes some issues and offers more flexibility, for example, in terms of spatial scale. Since data are based on administrative sources they are also usually more up to date and flexible. The first part of the paper describes the measurement of risk and methods for assessing the statistical significance of results, and then how to tabulate risk and interpret results systematically. Some local data sources are described and issues arising using data in specific situations are analysed. A case study on domestic violence then follows in which data are combined from several agencies to determine household categories most at risk. Results indicate a clear risk gradient depending on which factors are present or absent. Consideration is then given on how the risk model can be used for predictive purposes. A concluding discussion suggests how the techniques could be extended and used in a range of ways. GIS techniques are used to illustrate results.

1. Background and introduction

The cost of health care continues to increase, as demand seems to rise inexorably. Whilst the need for better health care services is undeniable, a substantial amount of need is socially determined through interaction between the social and health economy. Evidence for this is apparent in differences in life expectancy between economically deprived and more affluent areas, between the usage of services and wider social phenomena such as crime rates, domestic violence, drug abuse, unemployment and so forth. The Acheson report on inequalities in health (Acheson, 1998), for example, brings together the latest thinking on the causes of inequalities and provides many examples in a similar vein. However, a key problem is how to take this evidence down to the level at which appropriate policies can be devised and evaluated. This is because typically available data cannot show how localized these interactions are or indeed to what extent they occur within individual households. They may even lead one to false inferences or associations.

To give an example, suppose there is a high correlation at electoral ward level between domestic violence and household poverty. To infer the same correlation applies at individual household level risks the danger of committing what statisticians call the 'ecological fallacy' – that is drawing inferences from aggregate data about individual behaviour (Robinson, 1960; Greenland and Robins, 1994). However, inferences will also be sensitive to the size and number of zones, in this case wards, giving rise to what is generally referred to as the 'modifiable areal unit problem' and often considered as another form of 'ecological fallacy' (Openshaw and Taylor, 1981). Aside from this there are often practical problems as statistical units may not be co-terminus or change. A current example of the latter is electoral ward boundaries, recent changes to which have

caused havoc in the research community. In this paper, we side-step these issues by accessing household data directly from administrative systems and are therefore able to deal with recent data at a level of aggregation of our choosing.

At the conceptual level it is common ground that local services are often specialized and so professionals' knowledge is not easily shared or communicated to other agencies with a shared interest. This can lead to overlaps, lapses of communication or even worse well-publicized systemic failures, which in an ideal world it would be better to avoid. We call the interaction between the social environment and health care system and its impact on the health and well-being of the population the 'neighbourhood health economy' (see Figure 1). In effect we are saying that everything is connected in some way to everything else, and it is essentially a matter of sorting out the strength of different risk factors or combinations. A typical research question, for example, is to ask whether some households are more 'at risk' in terms of crime or adverse health events than others and to what extent such 'inequalities' make a difference to social and physical well being. In this paper we seek to quantify those risks in way that they can be compared systematically with other neighbourhoods, and so potentially help raise awareness among providers and to target resources more effectively.

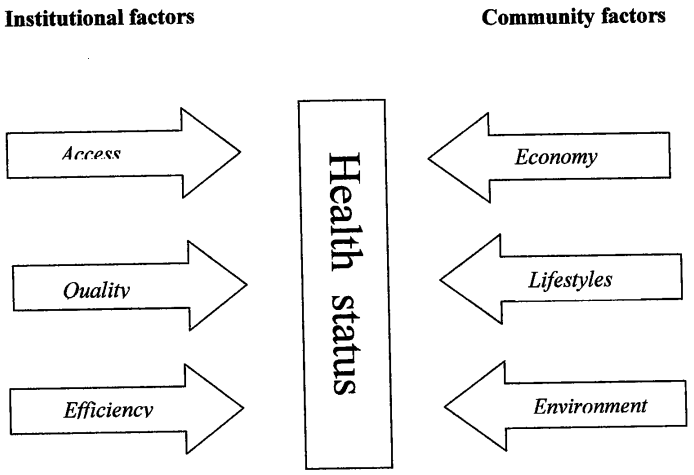


Figure 1: The Neighbourhood Health Economy showing how health status is determined not only by interventions of the health care system but by economic, lifestyle and environmental factors.

This paper arose from a project, financed by a grant from the Brent and Harrow Health Authority under the aegis of the Brent Health Action Zone (HAZ). HAZs, of which there

are some 26 across the country, can be regarded as the Labour Government's response to improving the way local services are delivered and health and social problems are addressed. Brent, which is located in northwest London, is often portrayed as an outer suburb but with inner city problems, and thus provides in many ways an ideal testing ground for the approach described here. The paper begins with a detailed examination of risk measurement before describing an illustrative case study on the subject of domestic violence, which is one of several undertaken in a similar vein. In more recent work, however, it has become clear that the methodology has other potential applications and these are briefly discussed at the end of the paper.

Whilst the methodology is built around basic statistical techniques and concepts, a key feature is the combined use of Geographical Information Systems (GIS) for holding and displaying spatial information as maps, which are used to analyze patterns and illustrate findings in some considerable detail. However, it is important not to trespass into areas of privacy, or breaches of confidentiality leading to inappropriate actions at the client interface and so for this reason the advice of the Data Protection Agency (DPA) was sought. Their advice was that personal data can be processed under section 33 of the 1998 Data Protection Act for research purposes, providing personal data are not disclosed or the results used to support decisions about individuals. Since the research would be used ultimately to improve the delivery of local services, this would not fall outside the expectation of local data subjects and so, on DPA advice, separate notification would not be necessary. This guidance was followed throughout the research.

The structure of the paper is as follows: Section 2 describes the statistical methodology, which is based around the concept of 'risk'. Section 3 describes excerpts from a domestic violence case study and further extensions of the methodology to include more general and predictive models, and section 4 discusses some research issues and extended applications.

2. Measuring risk

In this section, we outline the statistical methods used to evaluate probability and risk. Conceptually, the terms risk and probability are often used interchangeably and amount to the same thing technically speaking. We use the term risk here to define an adverse or disadvantageous event like bereavement or becoming unemployed, whereas we use the term probability in a more categorical way, for example the probability of living in social housing. However, these distinctions are not always hard and fast. Further technical detail of the mainly basic statistical techniques used may be found in statistical textbooks, for example Chou (1972), Barnett (1984) or Armitage and Berry (1987).

The methodology operates in two stages: firstly relevant data from different agencies are matched at address level using address matching techniques and the local authority property database (see Annex A for description of how this works). Each household is assigned a geographic co-ordinate and the data are then anonymized to suppress any identification of particular households. Secondly, individual factor combinations are tabulated and exhaustively enumerated and then analyzed for risk patterns.

Consider the following example, assuming three factors: whether 1) children in a household receive free school meals (FSM); 2) the household is based in public accommodation (council or housing association property), which we will subsequently describe generically as social housing; and 3) whether there has been a recent check up visit by the social services department. FSM indicate the presence of at least one child of school age and that a household is receiving Income Support, a means-tested benefit for households below the poverty line. Studies show social housing is more likely to be associated with certain socio-economic problems than owner occupied or privately rented households, whilst a visit from social services indicates a specific problem requiring their attention. There are eight possible combinations of factors as shown in Table 1.

No factors	One factor	Two factors	Three factors
1. None	2. Free school meals 3. Social services visit 4. Social housing	5. Free school meals & social housing 6. Social services visit & social housing 7. Free school meals and social services visit	8. Free school meals & social housing & social services visit

Table 1: Example of possible factor combinations in simple three-factor case

Suppose we are interested in the well being of children and suppose a health visitor is planning to contact certain households in an area. The relative frequency of each factor combination could indicate:

a. The geographic concentration of households with different factor combinations and the need for multi-agency co-ordination and referral.

b. The probability (i.e. likelihood) of finding a household where there may be child poverty or social problems.

Ideally it is desirable to work with as many factors as possible but there are two limitations (a) the availability of suitable data, and (b) the fact that the number of possible factor combinations grows very quickly. So for example with five factors the number of combinations rises to 32 while for 9 factors it increases to 512. The general mathematical rule is 2^N where N is the number of factors. One of these combinations has no factors present and is called the null set. In practice using various administrative data sets from the health and local authority and others we have found it is relatively unusual to find more than three factors present in any one household. However, households with more than three factors may still be of interest. In practice, this means we can usually limit our

investigation to no more than, say, five possible combinations from any number of factors. To enumerate combinations from N factors where m ranges from 0 to 5 we use:

$$\sum_{m=0}^{m=5} \frac{N!}{m!(N-m)!}.$$

A comparison with the previous formula shows the following differences for different values of N:

Number of factors	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
2^N	2	4	8	16	32	64	128	256	512	1024
$\sum_{m=0}^{m=5} \frac{N!}{m!(N-m)!}$	2	4	8	16	32	63	120	219	382	638

How is risk defined? Suppose there are 100 households in public or social ownership, all of which have received at least one visit from social services in a defined period and of these households 30 are entitled FSM. Suppose entitlement to FSM is used as a definition of deprivation because it signals entitlement to means tested Income Support. The risk that a household is 'deprived' is defined as 30/100 or 30% based on this sample of households.

We call this first example the *conditional probability or risk* because it is contingent on living in social housing. Suppose there are a further 50 households that have been visited by social services but in owner-occupied dwellings and of these 10 receive FSM. The risk is now 20% (10/50) whilst the *relative risk* of FSM in owner-occupied dwellings compared with social ownership is 0.67 (20%/30%). The *unconditional* risk of deprivation, or put another way the probability of social services visiting *any* household in receipt of FSM regardless of ownership, is defined as (30+10)/(100+50) or 26.7%.

Note risks are *asymmetric* depending on what one is attempting to measure. Suppose there are another 60 households receiving FSM that have *not* received a visit from social services. The probability of any household receiving FSM also receiving a visit from social services is (30+10)/(30+10+60) or 40%, which is not the same as the probability of a social services visiting a household receiving FSM. Statistically such distinctions are the same as writing the probability of A given B, Pr(A|B) or Pr(B|A). Later in the paper we also work with the concept of 'odds', but first we need to extend the analysis of risk.

Detecting unusual combinations or risks

It is of interest to know if the risks we measure are important and statistically significant and so we need to devise appropriate tests. If the number of households in a particular combination were unusual it might be because the factors are associated or attracted to each other. To check if the number in a combination is unusual we compare the expected number that would have arisen by pure chance with the number that actually occur.

Assume there are 100 households in an area of which 30 are social housing and the remainder owner occupied. Assume also 20 of the 30 receive FSM as well as 5 owner-occupied households. The remaining 65 households are neither social housing nor in receipt of FSM. The resultant sample space is represented in Figure 2 in which each dot is assumed to equate to 5 households.

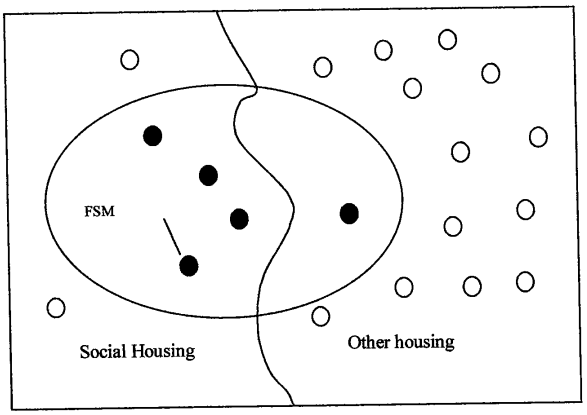


Figure 2: Sample space for households in receipt of FSM according to housing tenure. Each dot represents 5 households (see example)

We can calculate the expected probabilities and hence households using standard probabilistic methods assuming independence between the variables. For example the probability of finding FSM in owner occupied housing is:

$$p(FSM,\overline{SH}) = p(FSM)\,p(\overline{SH}) = 0.25x0.7 = 0.175$$

where the bar over SH indicates *not* social housing. Table 2 is a contingency table showing the observed and expected number of households in each category.

Observed	*FSM*	*Not FSM*	**Expected**	*FSM*	*Not FSM*
SH	20	10	*SH*	7.5	22.5
Not SH	5	65	*Not SH*	17.5	52.5

Table 2: Contingency table comparing observed and expected factor combinations

We use chi-squared statistic (χ^2) to test the null hypothesis that levels of risk are the same regardless of housing type. This yields:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(20-7.5)^2}{7.5} + \frac{(10-22.5)^2}{22.5} + \frac{(5-17.5)^2}{17.5} + \frac{(65-52.5)^2}{52.5} = 39.6$$

The critical value of χ^2 for 1 degree of freedom at the 5% level is 3.84 and so the null hypothesis that access to FSM is independent of housing tenancy can be safely rejected. The observed and expected frequencies in Table 2 can be re-expressed in terms of actual and expected 'risk'. Table 3 shows the risk of social housing receiving FSM is appreciably higher than the risk associated with chance. Similarly, it may be verified that the risk of receiving FSM in other housing is substantially less than the expected risk.

	Social housing receiving FSM	*Social housing not receiving FSM*	*Risk*
Observed (o)	20	10	$\frac{100x20}{20+10} = 66.7\%$
Expected (e)	7.5	22.5	$\frac{100x7.5}{22.5+7.5} = 25\%$

Table 3: Comparing actual and expected risks

It is also useful to give specific meaning to the concept of relative risk (RR) within a household receiving FSM but depending on household type. From Table 2 we can calculate this as the risk of SH receiving FSM divided by the risk of *not* SH receiving FSM or 66.7/7.1. This gives the relative risk value as being 9.33 times higher.

The above is extendable to situations in which there are multiple risk factors. If we do this the *expected* risk in any category will be same for the reason the factors are assumed to be independent of each other. Suppose there are n factors, the risk of occurrence of factor x within an arbitrary combination, say involving all n is:

$$r = \frac{p_1 p_2 \cdots p_{n-1} p_x}{p_1 p_2 \cdots p_{n-1} p_x + p_1 p_2 \cdots p_{n-1}(1-p_x)} = p_x$$

This result states that expected risk is a constant for any single factor regardless of other factors with which it is combined.

Whether an observed risk is higher or low than expected could influence service providers and policy makers in different ways but a risk is still a risk regardless of value. We need to systematize and structure how risk is measured and expressed in more complex cases with large numbers of factors, and this is the subject of following sections. Before we do this, however, we also need to have confidence the risk measured in this way is robust and reliable. Confidence intervals in this paper are expressed as the risk value plus or minus a range in which the true risk has a 95% certainty of occurrence. These are calculated using the following normal approximation to the binomial distribution, which is appropriate for use in the case of proportions (e.g. see Barnett, 1984, p42):

$$r \pm [z_\alpha \sqrt{\{(1-f)r(1-r)/(n-1)\}}\ ]$$

Here r is the measure of the risk of an event occurring and $1\text{-}r$ of it not occurring, n is the total sub-set of households over which risk r is being evaluated; z_α is the double-tailed value on the normal distribution assuming a 95% level of confidence; and f is the sampling fraction. In practice we assume f is small and can be neglected.

Consider the previous example. The risk of FSM and social housing is 66.7% (probability 0.67), the double tailed value of z at the 5 % level of probability is 1.96, and the sample size n is 30 (i.e. the number social housing households). These values yield an interval of 49.6% to 83.8% with a 95% level of confidence. In general a confidence range is narrower the greater the number of households in any given risk category, and so if the sample in this case were doubled to 60 the confidence interval would narrow to ± 12%.

Risks with a small probability of occurrence are harder to detect with confidence than larger risks, as seems intuitive. For example, if the risk is 5% and the sample is 100 the confidence interval would be ±85% of the risk estimate but if the sample size were 1000 it would be ±27%. Conversely if the risk were much higher, 20% say, and the sample size 100 the confidence interval would be ±39% of the risk estimate.

In typical case studies, sample sizes can vary from under 10 to over 100,000 depending on the category of household being considered and the risk being measured and so the problem of detecting small risks can be overcome to a degree. Occasionally very high values of risk are obtained (up to 100%) for small samples (below 5, say). These must be viewed cautiously since the normal approximation to the binomial distribution is no longer valid and the most one can say with confidence is that the risk is greater than 40% with 100% the upper bound (Table VI, p 483 , Freund 1973).

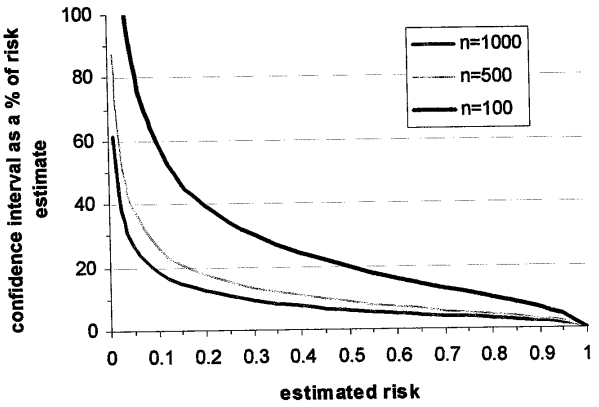


Figure 3: The effect of sample size on risk confidence interval expressed as a percentage of the risk estimate.

Figure 3 illustrates the relationship between confidence intervals and risk estimates and the impact of higher sample sizes using the normal approximation formula above. A simple criterion has been devised to assess the robustness of risk estimates based on this approach. If the confidence interval is within 10% of the estimated risk value three stars are assigned (***), 11%-20% (**), 21%-30% (*), and over 30% no stars.

The logic can be easily reversed in answer to the question of what sample size or number of households is required to achieve a given level of precision. This is given by the following formula:

$$n = \frac{z_\alpha^2 \hat{r}(1-\hat{r})}{x^2 \hat{r}^2} + 1$$

where $x \leq 1$ is the required level of precision and $\hat{r}$ is a prior estimate for a particular risk. Suppose the target was 10% precision so that $x = 0.1$ in this case and let $z = 1.96$ as before. If the prior expectation, $\hat{r}$ were 1% the sample needed would be 3804 but if it were 5% the sample required would be 730.

Confidence intervals can be similarly constructed for measures of relative risk. Consider two household categories as before, both receiving FSM. Assume the general layout of the of the *2x2* table is:

	FSM	*noFSM*	
SH	a	b	a+b
Not SH	c	d	c+d
total	a+c	b+d	n

The relative risk is given by $\frac{a}{a+c} / \frac{b}{b+d}$ and the standard error of the log of RR is given in Altman (1999) p266-268 as:

$$SE(\log_e RR) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$$

The sampling distribution approximates the normal distribution and so the 90 % confidence interval for the log of RR is:

$\log_e RR - N_{0.95} \underline{x}\, SE(\log_e RR)$ *to* $\log_e + N_{0.95} x\, SE(\log_e RR)$

Take the example in Table 2 where the RR is 9.33; after taking anti-logs the lower and upper 90% confidence intervals using the above formula are 4.5 and 19.6.

To the extent that risk varies by neighbourhhood one can construct spatially varying risk estimates by geographical location although sample sizes may be restrictive depending on the risk factor combination of interest. Beyond that there is theoretical scope to produce contour maps of risk, which would enable analyses of spatially varying risk patterns although whether such maps would be meaningful would depend on size of risk, sample size and so forth. The experience of this work however is these tests would rarely be met unless the samples were very large although maps constructed this way can be visually very effective.

Consider the diagrams in Figure 4 (a-c). The box represents the sample space of households in an area and the circles the households that are either social housing or FSM. In 1a it is seen FSM have a high affinity for social housing in 1b some affinity and in 1c negative affinity. Generally, most actual examples will fall into case (b) unless there are good reasons. For example, entitlement to FSM depends on households receiving Income Support so the FSM circle would be entirely contained within the Income Support circle in this case. Similarly, a crime victim is hardly likely to be the accused and so the victim and accused circles would be entirely separate as represented in case (c). Note for cases with inherent dependencies like (a) and (c) it does not make any sense to construct expected values of overlaps. These concepts are readily expanded to more detailed examples containing more than just two factors.

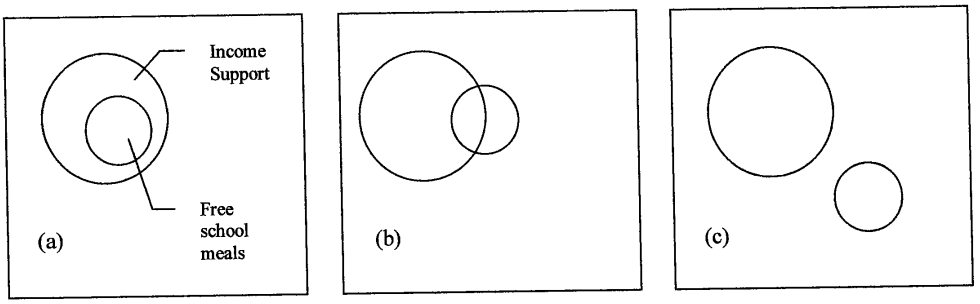


Figure 4: Association between FSM and social housing: (a) high affinity, (b) some affinity, (c) negative affinity

The relative size of overlaps and what to expect from data

The combining power of factors in any data set could be important in a large number of applications if the overlaps represent a cluster of social issues requiring targeted assistance. The relative number of households with overlapping factors will depend on the average incidence of each factor (occurrences per household), on systematic associations between each of the factors, and on chance. It is difficult therefore to give any general rules but a general indication based on expectation is possible if each factor is statistically independent of every other factor. However, it must be born in mind where there are systematic associations between factor risk structures will be more complex than is implied here.

Consider a neighbourhood in which five factors operate each with an equal chance of occurrence ranging from zero to one. Figure 5 shows the resulting composition of households. On the vertical axis is the expected proportion of households in which: no factors apply at all, 1 factor only, or 2 or more factors. As is seen the proportion of overlapping factors increases with the incidence of each factor while the proportion with no factors declines. The proportion with only one factor rises initially reaching a maximum when factor probability equals 0.2 but then falls again as multiple overlaps take control. When the incidence or average factor probability equals one all factors overlap with every household and so the proportion tends to one.

In general, a maximum expected proportion of single factor households occurs when the probability of occurrence equals *1/n,* where n is the number of factors. The peak arises when $P = (1-\frac{1}{n})^{n-1}$ as is easily shown. When n equals 2 it is 50% and 5, 40.1%. As n increases indefinitely the proportion converges to e^{-1} or 36.79% and the proportion of households with no factors converges to the same. Finally the proportion with 2 or more factors converges to $1-\frac{2}{e}$ which equates to 26.42%. However, these results are idiosyncratic since in reality dependencies among risk factors are more likely than not but they are useful to the extent they point the way to a more structured way of thinking about risk to which attention now turns.

Suppose we choose A as the risk factor, Table 5 shows how risk is tabulated in this representative example. Each combination provides the necessary basis for measuring the risk of A against a particular factor combination. It is evident the rows proceed systematically, first the risk of A with no other factor present, then with one factor, then with two and so on. These are the basic levels of possible risk combinations in a population. From the earlier example we saw that when one risk is expressed as a ratio of another it is called the *relative risk*. Table 6 shows the matrix of relative risks for the two-factor example in Table 5. An N factor model generates N^2 measures of relative risk in which the diagonal takes values of 1.

However, other definitions of risk flow from this table, which can be arranged in different levels to form a hierarchy and this is shown in Table 7. The symbol '$\cap$' means occurring together e.g. $A \cap B$ means the set of households in which factor A and factor B are both present whilst $A \cup B$ means 'either or'. The symbol Ω is defined as the universal set of all households. Column three shows how they are enumerated and should be read in conjunction with Figure 6. Each level represents different subsets or aggregations of risk.

Case	*B C*	*A*	$\bar{A}$	*Risk of A*
1	0 0	n_2	n_1	$\frac{n_2}{n_1 + n_2}$
2	1 0	n_5	n_3	$\frac{n_5}{n_5 + n_3}$
3	0 1	n_6	n_4	$\frac{n_6}{n_6 + n_4}$
4	1 1	n_8	n_7	$\frac{n_8}{n_8 + n_7}$
5		$\sum_{i \in A} n_i$	$\sum_{i \in \bar{A}} n_i$	$\frac{\sum_{i \in A} n_i}{\sum_{i \in A} n_i + \sum_{i \in \bar{A}} n_i}$

Table 5: Risk enumeration in the three-factor case.

case	AB	*00*	*10*	*01*	*11*
1	*00*	1	$\frac{r_1}{r_2}$	$\frac{r_1}{r_3}$	$\frac{r_1}{r_4}$
2	*10*	$\frac{r_2}{r_1}$	1	$\frac{r_2}{r_3}$	$\frac{r_2}{r_4}$
3	*01*	$\frac{r_3}{r_1}$	$\frac{r_3}{r_2}$	1	$\frac{r_3}{r_4}$
4	*11*	$\frac{r_4}{r_1}$	$\frac{r_4}{r_2}$	$\frac{r_4}{r_3}$	1

Table 6: Table of relative risk based on a two-factor model

There are two risk sets represented. The *conditional set* shows risk combinations built on component subsets factors having regard to their presence or absence. The *unconditional set* is based on factor combinations in which one or more factors are ignored for calculation purposes and can be regarded as a form of aggregation.
In more detail:

- *Level 0* in Table 7 is associated with single combinations, e.g. A on its own (conditional set) or A occurring at all (unconditional set).

- *Level 1* is a measure of the risk of A combining with B with C absent (conditional) or the risk of A combining with B regardless of whether C present (unconditional).

- *Level 2* is a measure of A combining with factor C with B absent (main set) or of A combining with C regardless of whether B present.

- *Level 4* is the risk of A occurring with B *and* C (conditional) or the risk of A occurring with B *or* C (unconditional).

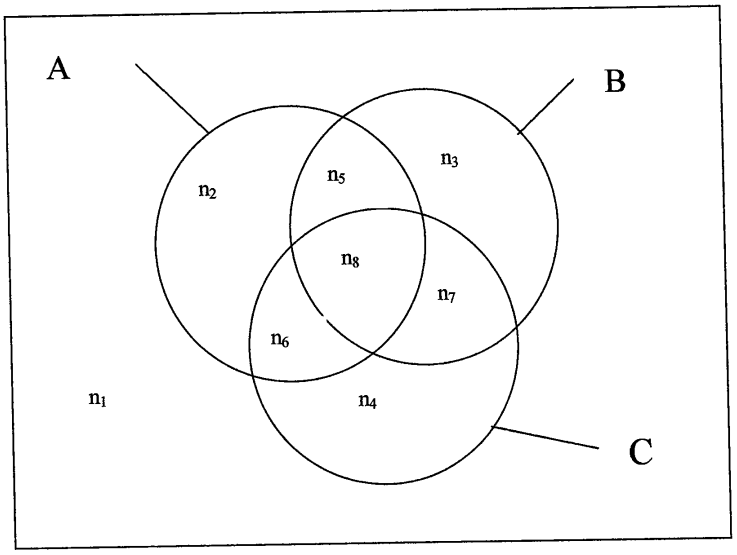


Figure 6: Counting units for evaluating risk. See also tables 4 and 5.

Level	*Risk set*	*Evaluation*	*Conditional risk or probability*
Conditional set			
0	$\frac{A \cap \bar{B} \cap \bar{C}}{\bar{B} \cap \bar{C}}$	$\frac{n_2}{n_2 + n_1}$	Risk of A occurring on its own
1	$\frac{A \cap B \cap \bar{C}}{B \cap \bar{C}}$	$\frac{n_5}{n_5 + n_3}$	Risk of A occurring with B with C absent
	$\frac{A \cap \bar{B} \cap C}{\bar{B} \cap C}$	$\frac{n_6}{n_6 + n_4}$	Risk of A occurring with C with B absent
2	$\frac{A \cap B \cap C}{B \cap C}$	$\frac{n_8}{n_8 + n_7}$	Risk of A occurring with B and C
Un-conditional set			
0	$\frac{A}{\Omega}$	$\frac{n_2 + n_5 + n_6 + n_8}{n_2 + n_5 + n_6 + n_8 + n_1 + n_3 + n_4 + n_7}$	Risk of A occurring at all
1	$\frac{A \cap B}{B}$	$\frac{n_5 + n_8}{n_5 + n_8 + n_3 + n_7}$	Risk of A occurring with B regardless of whether C present
	$\frac{A \cap C}{C}$	$\frac{n_6 + n_8}{n_6 + n_8 + n_4 + n_7}$	Risk of A occurring with C regardless of whether B present
2	$\frac{A \cap (B \cup C)}{B \cup C}$	$\frac{n_5 + n_6 + n_8}{n_5 + n_6 + n_8 + n_3 + n_4 + n_7}$	Risk of A occurring with B or C

Table 7: Levels of risk in three-factor case

Enumerating risk hierarchies

The number of risk estimates therefore depends on the number of factors. In the example given above there are three levels of risk, two sets and eight separate risk measures. Table 8 shows the risk combinations based on four factors arranged in a hierarchy. In general the number of levels and risk measurements follow coefficients of a binomial expansion as follows:

No. of factors					
1			1	1	
2			1	2	1
3		1	3	3	1
4	1	4	6	4	1

Conditional set				
		1100		
		1010		
	1000	1001	1110	
	0100	0110	1101	
	0010	0101	1011	
0000	0001	0011	0111	1111
level 0	*level 1*	*level 2*	*level 3*	*level 4*
****	***1	**11	*111	1111
	**1*	*1*1	1*11	
	*1**	*11*	11*1	
	1***	1**1	111*	
		11**		
		1*1*		
Unconditional set				

Table 8: Risk hierarchy for four factors for the conditional and unconditional risk sets. Zero indicates absence of a factor and an asterisk indicates factor is ignored.

Data considerations

Thus far we have only indicated the types of data that may be used but have not provided any detail of how they are selected or processed. Here we are concerned with prior selection considerations, which to some degree may be dictated by the form in which data are held. Essentially, however, there are three types of data, which may loosely be described as 'categorical', 'event', or 'flow'.

Categorical data are data that do not change very often or at all, and are sometimes called 'state' variables. Examples would include household tenancy or the ethnicity of a subject. Event data are data of the one-off variety like a birth, bereavement or crime, whereas flow data refer to things with a measurable duration. This could include spells on Income Support or entitlement to FSM for which the household base changes quite frequently as people move in and out of work. The three types of data are fundamentally different in the sense that the stock of social housing at a point in time does not alter over the short term. The stock of households receiving FSM, by contrast, is related to the flow (incidence) and the duration of entitlement (in the steady state, stock = flow x average duration). The stock of 'events' however is one or zero if one accepts that two events cannot occur at precisely the same time.

We used primary sources of data that ranged from publicly accessible registers of births, deaths and marriages to administrative records held by organizations ranging from local health, education and police. The simplest way of associating data is to pick the most recent time period, say two years and to match occurrences over that time frame, the hypothesis being factors are 'associated' if they occur together in the same household. Whilst this appears straightforward in principle, there are a number of important issues to consider especially with 'flow' and 'event' data types.

Administrative data for example tend to reflect the current caseload and not lapsed cases and access to lapsed cases that terminate within the chosen time frame may not be possible or least extremely difficult. This inevitably introduces an element of imprecision into calculations involving this type of data and also a source of bias if lapsed cases are not typical. Also some subjects change address during the interval and so working with 'households' rather than named individuals can be another source of inaccuracy. However, these problems are potentially surmountable as data sources are improved.

With event data the issue is more to do with dealing with multiple occurrences of events at the same address. A typical example of this would multiple visits from a health worker or several reported crimes at the same address. One option is to count multiple events as single occurrences or alternatively each event could be counted separately. A generic problem occurs where it is not mandatory to report something to an authority and this will lead potential to gaps in the analysis. A classic example of this is crime and the fact not all crimes or relevant incidents may be reported can introduce bias or under estimation of the risk or risks involved. A health service example arises where for reasons of resource constraints the methodology may not be directly translatable into need.

There are several other aspects to typically available data that can limit the depth or accuracy of analysis or prevent its use altogether. The first is using data from voluntary organizations that are incomplete in some way, or of small sample size, or limited to small sub-groups of the population. Such data sets obviously need to be used with caution and in fact often turn out to be useless. An important example occurs with respect to the availability of ethnic data, which is typically included in some data sets such as reported crime but not for others, such as household tenancy. In this case, it may only be feasible

to evaluate risk in certain factor combinations and at certain levels in the risk hierarchy. However, such analyses may be misleading.

Take for example the case of burglary. Crime data record the ethnic group of victims and accused and their addresses, but suppose there is no information on the ethnicity of households *not* victimized or accused of burglary. In such situations it would be possible to analyze the risk of being victimized or accused by ethnic group, while controlling for other factors like household tenancy and so on to see which are more susceptible. However, such an analysis could not provide an overall risk assessment of the likelihood of a particular ethnic group being involved in that crime unless their total numbers can be enumerated through other data. We may illustrate this by returning to the three-factor example in Figure 6 and the hierarchy in Table 7. Suppose factor A represented suspected burglars, factor B an ethnic group, and C social housing. Estimates would not normally be available for n_3, n_7, n_4, or n_1 in Figure 6. These represent:

- n_3 the number of households in the chosen ethnic group *not* involved in burglary and *not* living in social housing,

- n_7 the number in the chosen ethnic group that *are* living in social housing but *not* involved in burglary,

- n_4 the number *not* in the ethnic group that *are* living in social housing but *not* involved in burglary and

- n_1 the number *not* in the ethnic group *not* living in social housing and *not* involved in burglary.

Note however, we will generally have or can deduce information on n_4+n_7 or n_3+n_1. More to the point, if we had information on just one of the missing unknowns all the others could be calculated. In practical terms this means *it may not be possible to calculate risk in certain levels or combinations.* In this particular example only two can be imputed from the list in Table 4 with the information available. These are:

- Level 0, unconditional set, the risk of A occurring at all (i.e. the overall risk of burglary)

- Level 1, unconditional set, the risk of A occurring with C regardless of whether B is present (i.e. burglary with social housing), and

Neither provides an ethnic based risk assessment. If such data are missing some of the possible analyses that can be carried out will be misleading. For example, if we consider all burglaries only (circle A in Figure 6) and examine the ethnic composition of the accused we might find one group over represented, but this may not be a surprising result if that group represent a majority of all households. Thus, it is preferable to have information about all households not just those accused or suspected of burglary.

Data Set	*Data Type*	*Source*
Property database *-residences* *-tenancy*	Categorical	Local authority
Free school meals	Flow	Local education authority
Home health visits	Event	Community health trust
Home dietician visits	Event	Community health trust
Home physiotherapy visits	Event	Community health trust
Home SLT[1] visits	Event	Community health trust
Registered mental health Patients	Flow	Mental health trust
Bereavements	Event	Register of births deaths and marriages
Births Register	Event	Register of births deaths and marriages
Noise complaints	Event	Local authority
Reported domestic violence: - *Victims* - *Accused/suspected*	Event	Police authority
Reported teenage crime	Event	Police authority
School exclusions	Event	Local education authority
School pupil roll	Flow	Local education authority
Vulnerable persons register	Flow	Local authority
Brent Housing priority list	Flow	Local authority

Table 9: Data sets featured in recent case studies.[1] Speech and learning therapy.

Table 9 is a list of administrative and publicly accessible data sets we have used in recent case studies. Not all administrative systems provide data at address level to enable matching to the property database. Hospital activity data, for example, only records post-code. The ambulance service, on the other hand, provides a locational reference that is inconsistent with the referencing system used in the local authority property database. One obvious reason for this is that accidents do not necessarily take place in the home but elsewhere. Another shortcoming is local data about children especially the number in a household, but one possible source, the electoral register, only records adults. An emergent source about children is school pupil rolls, which are currently being centralized at local authority level. Such developments lead inevitably to the conclusion that the number of data sources will expand greatly in coming years rendering all kinds of linkages possible in the future.

3. Illustrative case study

In this section we reflect on the results of a case study on the subject of domestic violence (DV), which is based on the risk framework described and is one of several undertaken over the past two years. Domestic violence (DV) is defined in our case study in terms of incidents reported to the police over an 18-month period. Police classify incidents as 'domestic' if they involve any member or member of a household including children or there is some sort of close relationship, for example a former partner. Figure 7 is a contour map of DV in Brent based on incidents per hectare over an 18-month period and clearly shows there are marked concentrations in certain areas. Our aim in this illustration is to ascertain whether there are any systematic risk variations based on five other factors: 1) households in receipt of free school meals (FSM); 2) households with registered mental health patients; 3) noise complaints; 4) drug offending and 5) social housing (council tenancy or housing association).

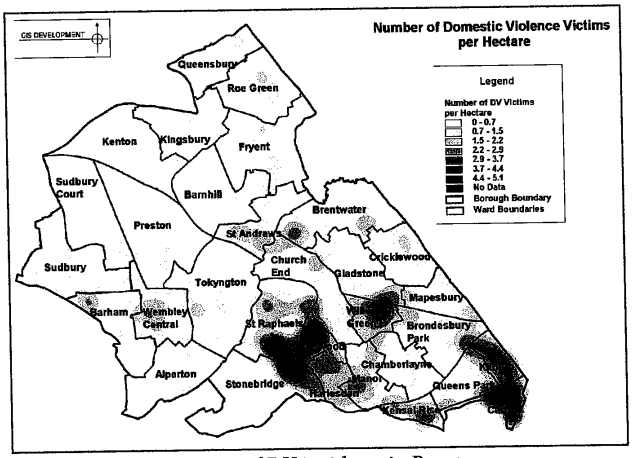


Figure 7: Map showing the density of DV incidents in Brent.

The total number of factor combinations again totals 64 giving rise to 32 risk categories in the conditional set and another 32 in the unconditional set. Note the threshold for the DV is set high since it depends on an incident being reported to the police and so that true levels of DV (reported and unreported) are bound to be higher. How risk varies in this more general case is beyond the scope of this paper since it would require more detailed data and surveys of households. Initially we count multiple reports of DV from the same household as a 'single incident' but later we deal with multiple incidents separately. Altogether four cases are considered: conditional and unconditional risk sets, and any incident and multiple incidents.

What should we be looking for in the results? First, we should satisfy ourselves that risk estimates differ significantly from the expected risk if all the factors were independent using techniques previously outlined. Secondly we should examine the hypothesis risk is higher when there are more risk factors but we also should identify whether individual risk estimates are in some sense consistent. For example, suppose the risk of DV in A and B is *w*, in A and C, *x* and B and C, *y*, is the risk of A and B and C, *z*, greater than *w*, *x*, and *y*? Thirdly, we should be interested in whether some factors have greater risk than others and if factors are independent of one another. One suitable test of independence would be if the addition of a factor had the same effect regardless of any other factors present.

Using complete samples within a defined period it is possible to answer these questions reasonably definitively through complete enumeration of all risks combinations. A problem occurs however in testing questions of independence and the relative strength of each factor without some prior expectation or benchmark. We used logistic regression for this purpose to provide a statistical 'fit' to and to quantify the relative strength of each risk factor, and also to investigate if there are any systematic dependencies between individual factors.

A description of logistic regression techniques can be found in many statistical textbooks and so it is sufficient to be brief here (e.g. Armitage and Berry, 1987). Let the risk of DV occurring be r . Define the odds of DV occurring as $\frac{r}{1-r}$ and assume that $\log(\frac{r}{1-r})$ is linearly related to the risk factors x_i, which take the value of 1 (factor present) or 0 (factor absent). The following equation:

$$\ln(DVodds) = \log(\frac{r}{1-r}) = \alpha + \sum_i \beta_i x_i + \varepsilon_i$$

is called a logit model in which logit estimates give information on the partial effect of each variable on the risk of DV in which alpha and beta are regression parameters and ε_i is the residual term. Ignoring ε for the moment and rearranging we have:

$$r = \frac{odds}{1 + odds}$$

Which is equivalent to:

$$r = \frac{\exp(\alpha + \sum_i \beta_i x_i)}{1 + \exp(\alpha + \sum_i \beta_i x_i)}$$

$Exp(\beta_i)$ is the odds of DV occurring in the presence of drug offending when all the other parameters are held fixed. In logistic regression the parameters are usually estimated using maximum likelihood techniques within specialized statistical package but another method of estimation giving slightly different results is using ordinary least squares (OLS) regression in a spreadsheet package. The data used for both methods are contained in Tables 9-12, which are based on actual risks for each fully enumerated factor combination for which there is at least one observation.

The key difference between the two approaches is that the OLS method is based on risk combinations and weights each combination equally regardless of the number of observations. The maximum likelihood method, on the other hand, treats each household individually and therefore uses all the information in the data set but there are also other methods including weighted least squares. However, a detailed comparison of various methods used and results obtained are beyond the scope of this paper. Suffice it to say our test of acceptability is based on the ability of the model to describe accurately the risk associated with each combination rather than as a whole. Whilst both maximum likelihood and OLS procedures produced consistent results, we found OLS provided a marginally better *overall* representation of based on plots of observed and predicted risk. This may well differ in other examples and so a pluralistic approach seems sensible.

The model assumes that each factor is independent of the others so that if two factors are present, say drug offending and FSM, the odds would simply be the product of two exponential terms. In practice it could be possible that actual behaviour is more complex. For example, the presence of one factor with another could cause second order or interaction effects raising risk (or reducing it) compared to where it would be if the factors were statistically independent. However, when this possibility was tested for all possible interaction pairs only weak effects were observed and so the detail is omitted for the sake of brevity.

Measuring the risk of any DV incident

Tables 10 and 11 shows factor combinations in descending order of risk for the conditional and unconditional risk sets treating every household the same regardless of the number of incidents reported. It is noteworthy that the unconditional set follows a

similar pattern to the conditional set although, by definition households can now be members of more than one risk category at a time. This means the column sum of households is meaningless and so interpretation of the results tends to be more difficult for this reason, but it is included for completeness. Risk levels are now either higher or the same for any given factor combination.

The first two columns show risk levels and households in each particular factor combination. Since there are five risk factors, levels can range up to five although in practice the maximum was 3 (higher level combinations being empty). If the number of households is 100 and the risk 5% it implies 1 in 20 households in this category are at risk of DV. The last column refers to the confidence interval around the estimate according to the convention discussed in a previous section. Results are tabulated in risk order from high to low and in general we see that the more factors there are, the higher the level of risk. On the other hand the number of households in high-risk categories is markedly smaller which was to be expected based on our earlier analysis. However, the fact that risks differ for each factor combination is itself an indication of dependencies in the data.

At highest risk are households in receipt of FSM and where there have been noise complaints. In Table 10 the risk turns out to be 100% although there are only two households in this category out of a total of 102,427 and is not statistically significant. The second risk category, FSM and drug offending, has a risk of 33% but here also there are only three observations. Moving down the table to factor combinations with more sizeable numbers of observations are those involving social housing and drug offending (11.8%), social housing and FSM (9.6%), FSM by itself (6.7%) and social housing by itself (3.8%). The largest category in terms of households (74,356) is that with no factors present (1.8%).

One can also obtain estimates of relative risk by dividing one risk by another. So for example the risk of DV in the presence of drug offending and social housing compared with the risk for FSM and social housing would be 11.8/9.6, which is 1.22 or 22% higher, with associated 90% confidence intervals of 0.92 to 1.63. The odds are given by $\frac{r_n}{100-r_n}$ if risk is expressed as a percentage so the relative odds of two risk categories n and m would be $\frac{r_n}{r_m}(\frac{100-r_m}{100-r_n})$ or 25% higher in this example. Annex B gives further details of relative risk in relation to empirical estimates for two of the following examples and presents examples of relative risk 'look up' tables for ease of reference. These enable one to consider the relative risk of any one particular combination of risk factors against another in a convenient and easy to use form. They can be constructed using either observed data, although this may lead to gaps if some risk groups are empty or using predicted risk based on the logistic model.

Risk level	No. of households	noise	mental health	social housing	drug offence	free school meals	% risk of DV	risk estimate
3	2	Y		Y		Y	100.0	
2	3		Y		Y		33.3	
2	19				Y	Y	26.3	
3	34			Y	Y	Y	20.6	
3	15		Y	Y		Y	13.3	
2	323			Y	Y		11.8	*
3	10		Y	Y	Y		10.0	
2	1265			Y		Y	9.6	* *
2	11		Y			Y	9.1	
1	405				Y		6.9	
2	366		Y	Y			6.8	
1	1209					Y	6.7	*
2	35	Y		Y			5.7	
1	374		Y				5.3	
1	23944			Y			3.8	* * *
0	74356						1.8	* * *
1	49	Y					0	
2	2	Y	Y				0	
2	3	Y				Y	0	
3	2	Y	Y	Y			0	

Table 10: Risk of DV incident being reported to the police according to different factor combinations in Brent based on 102,427 households and 31 mutually exclusive categories in descending order; 11 categories are omitted from the list as they contain no households and therefore no DV. Asterisks indicate the robustness of the risk estimate (see convention earlier).

There are some consistencies but also some inconsistencies between the risk categories. For example the risk of noise, social housing and FSM is greater than the risk of noise and social housing, and social housing and FSM, and noise and FSM. However, the evidence is inconclusive, as there are so few cases in either category. Conversely, if we take the combination of drug offences and FSM, the risk is nearly 6% higher than the risk category social housing, drug offending and FSM. However, this finding is probably an anomaly and the consequence of only having small samples in these particular groups. Four risk factor combinations have no record of DV whilst a further 11categories (not shown) had no households in them at all.

A key conclusion of this section, evident in both Tables 10 and 11, is that there is a risk gradient between different household categories. A chi-square test of the hypothesis that DV is independent of household category would be rejected for this reason as is easily demonstrated. Figure 8 shows this in a different way using a graph in which risk is plotted on the vertical axis and each factor combination on the horizontal axis using the data in Table 10. Vertical lines display the confidence intervals applying in each case, where obviously one is looking for narrow confidence bands wherever possible. The hatched horizontal line in the graph shows the average risk of DV in all 102,427 households (2.5%).

Risk level	No. of households	noise	mental health	social housing	drug offences	free school meals	% risk of DV	risk estimate
3	2	Y	-	Y	-	Y	100.0	
2	5	Y	-	-	-	Y	40.0	
2	53	-	-	-	Y	Y	22.6	
2	34	-	-	Y	Y	Y	20.6	
3	13	-	Y	-	Y	-	15.4	
3	15	-	Y	Y	-	Y	13.3	
2	367	-	-	Y	Y	-	12.5	*
2	26	-	Y	-	-	Y	11.5	
1	39	Y	-	Y	-	-	10.3	
3	1316	-	-	Y	-	Y	10.1	**
2	794	-	-	-	Y	-	10.1	**
2	10	-	Y	Y	Y	-	10.0	
1	2558	-	-	-	-	Y	8.6	**
2	393	-	Y	Y	-	-	7.1	
1	783	-	Y	-	-	-	6.4	*
1	93	Y	-	-	-	-	4.3	
1	25996	-	-	Y	-	-	4.2	***
0	102427	-	-	-	-	-	2.5	***
2	4	Y	-	Y	-	-	0	
3	2	Y	-	Y	Y	-	0	

Table 11: The unconditional risk set in which risk categories are no longer mutually exclusive and so their sum exceeds the total households in Brent. The table shows for example that the average risk of DV in Brent in the relevant time period, ignoring other factors, was 2.5% (third line from bottom) and the risk of DV in the presence of drug offences and any other factor is 10.1%.

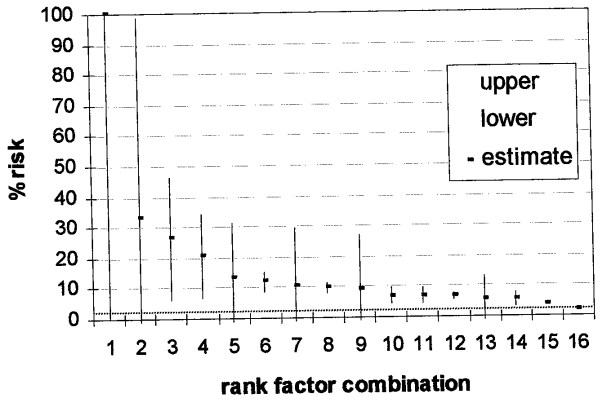


Figure 8: Factor combination risks taken from the main set and their confidence intervals as indicate by vertical bars.

Table 12 shows the results of the logistic regression for the conditional and unconditional cases in Tables 10 and 11 using the OLS method. Annex B gives results based on both OLS and ML including tables of observed and predicted values. The model parameters and overall goodness of fit based on R^2 are reasonable. Plots of observed and expected risk values (also shown in annex B) show that all but the extreme observations of over 33% based on small samples fall on a straight line with a gradient close to one. The results for the conditional case indicate that drug offending is the highest risk factor, increasing the odds of DV 3.9 times, followed by FSM (2.6), noise (2.0), mental health (2.0), and social housing (1.1).

It is noteworthy that the coefficient for social housing in the conditional case, which has an associated t-statistic of only 0.34, is not *of itself* significant at predicting the *level* of DV. A possible explanation for this is that social housing may be considered a categorical variable rather than an active variable, but one that is rather more important in determining the context in which DV occurs. Figure 9, for example, is a spider chart showing average incidence of each risk factor according to housing type. As is seen overall level of incidence of DV, FSM, mental health problems and noise is higher in social housing but the relative incidence of each factor is roughly the same. The main difference is that in social housing the problems appear to be magnified two to three fold. We conclude therefore that social housing is an important differentiating characteristic in this case and is more likely to occur in this housing setting.

	Model	constant term	noise	mental health	social housing	drug offending	free school meals	R^2
Conditional	β	-3.59	0.7	0.68	0.09	1.35	0.96	0.8
	s.e	0.29	0.56	0.26	0.25	0.26	0.26	
	t	-12.38	1.25	2.68	0.34	5.1	3.64	
Unconditional	β	-3.31	0.57	0.39	0.20	1.01	0.9	0.86
	s.e	0.18	0.26	0.16	0.14	0.16	0.16	
	t	-18.63	2.19	2.44	1.35	6.22	5.57	

Table 12: Logistic regression results for any DV incident – conditional and unconditional cases

Multiple incidents of domestic violence at the same address

In this section we consider whether the risk profile changes for households where more than one incident is reported. This could occur in more 'hardened' cases, perhaps because of the presence of one or more particular risk factors. Multiple occurrences are not uncommon and things to consider here include whether the overall rankings of risk categories are preserved. The results for the conditional and unconditional sets are reported in Tables 13 and 14. Generally, the results show the risk of multiple incidents falls on average by nearly two-thirds, suggesting that police reporting may have lead to a reduction in incidents. They show rank order is more or less preserved compared to before with the exception of one or two risk categories involving small numbers of households.

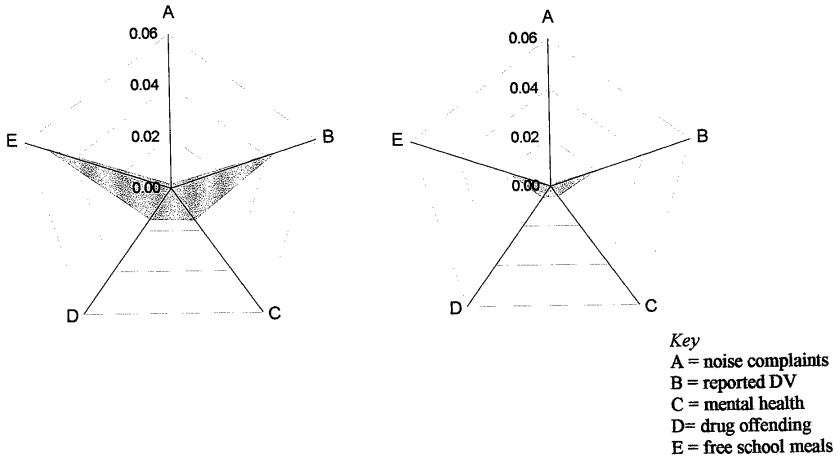


Figure 9: Spider diagrams showing the relative incidence of each factor split by social housing and other housing. Left shows social housing category and right other housing.

level	*No. of households*	*noise*	*mental health*	*social housing*	*drug offence*	*free school meals*	*% risk of multiple DV*	*risk estimate*
3	2	Y		Y		Y	100.0	
2	3		Y		Y		33.3	
3	34			Y	Y	Y	11.8	
2	19				Y	Y	10.5	
3	10		Y	Y	Y		10.0	
3	15		Y	Y		Y	6.7	
2	35	Y		Y			5.7	
2	323			Y	Y		5.3	
2	1265			Y		Y	4.1	*
2	366		Y	Y			3.0	
1	374		Y				2.9	
1	1209					Y	2.8	
1	405				Y		2.5	
1	23944			Y			1.4	***
0	74356						0.6	***
1	49	Y					0	
2	2	Y	Y				0	
2	3	Y				Y	0	
2	11		Y			Y	0	
3	2	Y	Y	Y			0	

Table 13: Risk of a DV incident being reported more than once to the police from the same household (the conditional set). 11 categories (not shown) have no households.

Level	No. of households	noise	mental health	social housing	drug offence	free school meals	% risk of multiple DV	risk estimate
3	2	Y	-	Y	-	Y	100.0	
2	5	Y	-	-	-	Y	40.0	
2	13	-	Y	-	Y	-	15.4	
3	34	-	-	Y	Y	Y	11.8	
2	53	-	-	-	Y	Y	11.3	
2	39	Y	-	Y	-	-	10.3	
3	10	-	Y	Y	Y	-	10.0	
3	15	-	Y	Y	-	Y	6.7	
2	367	-	-	Y	Y	-	6.0	
2	1316	-	-	Y	-	Y	4.5	*
1	794	-	-	-	Y	-	4.4	
1	93	Y	-	-	-	-	4.3	
2	26	-	Y	-	-	Y	3.8	
1	2558	-		-	-	Y	3.7	**
2	393	-	Y	Y	-	-	3.3	
1	783	-	Y		-	-	3.2	
1	25996	-	-	Y	-	-	1.6	***
0	102427	-	-	-	-	-	0.9	***
2	4	Y	Y	-	-	-	0	
3	2	Y	Y	Y	-	-	0	

Table 14: Risk of a DV incident being reported more than once to the police from the same household (the unconditional set). The risk for all Brent in the relevant time period for repeated DV incidents was 0.9% (third line up from bottom).

Table 14 shows the logistic regression results for the multiple incident cases for the conditional and unconditional risk sets. The main differences compared with the previous results are that noise, drug offending and mental health now present a somewhat higher risk. Based on the conditional set noise complaints increase the odds of DV 6.5 times compared with 2 times previously, drug offending 5.1 times (3.9 times), and mental health 3.7 times (2.0 times) and FSM 2.9 times (2.6 times). The results seem intuitively reasonable and it is expected that multiple reported cases are in some sense more hardened and difficult to stop than isolated incidents.

	Model	constant term	noise	mental health	social housing	drug offending	free school meals	R^2
Conditional	β	-4.71	1.86	1.32	0.04	1.62	1.07	0.84
	s.e	0.34	0.65	0.33	0.31	0.31	0.33	
	t	-13.84	2.86	3.96	0.13	5.23	3.22	
Unconditional	β	-4.48	2.18	0.79	0.17	1.48	1.07	0.85
	s.e	0.28	0.34	0.26	0.23	0.26	0.23	
	t	-16.10	6.32	3.08	0.75	5.76	4.57	

Table 15: Logistic regression results for the multiple incident case

Predicting reported DV incidents

It is important to understand not only the factors associated with a particular risk such as DV, but also whether such information has predictive value. A question for those interested in this field is by how much DV would rise or fall if one or more associated factors were to change, for example because of an improving economic situation. In principle, if the level of risk stays the same for a given factor combination it should be possible to estimate changes of this nature based on the dependencies between the factors and alterations in the sizes of the factor combinations.

One way to proceed might be to use simulation techniques to generate large number of household types with the appropriate factor combinations. However, such techniques normally assume factors are *independent* of each other and, whilst the results may appear reasonable (as may be easily demonstrated) they do not, clearly, reflect correctly the actual information in the 'risk ladders' e.g. Tables 10,11,13, or 14, which are based on information taken from all households. An alternative approach is to estimate the change in household composition directly, assuming that risk, in this case DV, remains the same.

To set the scene for the calculations required Table 16 shows a simplified 3-factor case in which a percentage change occurs in one of the factors, in this case (A). The approach involves two steps. Firstly, an estimate is made of the new number of households in each category consequent on a change in A, either a percentage increase or a decrease as given by *f* (see columns (6)-(9)). Secondly, based on the revised number of households in each category, the appropriate risk factor for that category is applied to obtain an estimate of the revised number of cases.

Col (1)	*Col(2)*	*Col(3)*	*Col(4)*	*Col(5)*	Col(6)	Col(7)	Col(8)	Col(9)
ABC	*Number of cases*	*A*	*B*	*C*	Factor increase		factor reduction	
					(-)	*(+)*	*(-)*	*(+)*
000	n_1	-	-	-	N_1f_1	-	-	n_2f_2
100	n_2	n_2	-	-	-	n_1f_1	n_2f_2	-
010	n_3	-	n_3	-	n_3f_1	-	-	n_5f_2
001	n_4	-	-	n_4	n_4f_1	-	-	n_6f_2
110	n_5	n_5	n_5	-	-	n_3f_1	n_5f_2	-
101	n_6	n_6	-	n_6	-	n_4f_1	n_6f_2	-
011	n_7	-	n_7	n_7	n_7f_1	-	-	n_8f_2
111	n_8	n_8	n_8	n_8	-	n_7f_1	n_7f_2	-

Table 16: Tabulating changes in factor combinations following a proportionate increase or reduction in one factor.

An issue arises where there are changes in several factors at once, since the results obtained are sensitive to the order in which the calculations are carried out. For example, the results will be different, if the effects of a change in B or C are considered before a change in A, but in the real world all such changes may be occurring simultaneously. This is an aspect that would be worth consideration in further development of the methodology and is not pursued here, although as far as the data used in this example are concerned the differences are not so sensitive as to invalidate the conclusions. For example, if the numerical example below were to be repeated by including factors sequentially the effects would not be materially different except perhaps at very high proportionate reductions.

Table 17 shows the result of systematically reducing the incidence each of the factors in the DV example by 0.1, 0.25, 0.5, 0.75 and 1, in which the effect of each factor is calculated independently of every other factor. The implicit assumption is that such changes would not be due to administrative changes, such as the re-designation of social housing or a change in the definition of mental health or the eligibility rules for FSM, but 'real' shifts changes in the circumstances of the population.

The results show that whilst the risk of DV is high in cases involving for example noise complaints, and drug offences the impact of a reduction in their incidence on the total number of reported DV cases will be relatively small. This is because there are far fewer incidents of noise complaints and drug offending than there are households receiving FSM which shows a much greater associated case reduction. It also shows that biggest reduction would occur in the social housing sector, which is numerically very large, but where the additional risks of DV are modest compared with other factors. It is also noteworthy that even if all the factors were reduced 100% total DV notification would reduce by 755 cases leaving 1827 cases. This suggests there are other factors, as yet unaccounted for in the model, which still need to be identified in order to explain underlying levels of DV.

	Expected case reduction in DV				
Proportionate reduction in given factor	*0.1*	*0.25*	*0.5*	*0.75*	*1*
Noise	0	0	0	1	1
mental health	3	6	13	19	26
Social housing	53	132	264	397	528
drug offences	6	14	28	41	55
FSM	14	36	72	108	144
total case reduction	*76*	*188*	*378*	*567*	*755*

Table 17: Predicted reductions in reported incidents of DV following given proportionate reductions in each risk factor.

The ecological fallacy in perspective

Because these results apply at household level we have effectively side stepped the danger of drawing erroneous conclusions from aggregate data. If we had undertaken this analysis in the traditional way by building a regression model on ward level data, what kind of results might we have obtained? Figure 10 is a plot of DV per hundred households versus FSM per hundred households in which each household is counted once if it has reported one or more incidents over the time period. The results demonstrate a high correlation whilst the best-fit equation could imply, at least on one interpretation, a majority of households receiving FSM have reported a domestic violence incident.

However, using household level data the percentage receiving FSM *and* reporting DV is only 8.4% suggesting whilst DV and FSM occur in the same *neighbourhood*, it would be wrong to infer they occur in the same *household*. Thus, although household level analysis shows no disagreement that FSM is a risk factor it is not as great as is suggested using regression analysis based on aggregate data. This result adds weight therefore to the need for caution in interpreting spatially aggregated statistical models and illustrates another advantage of using household level data.

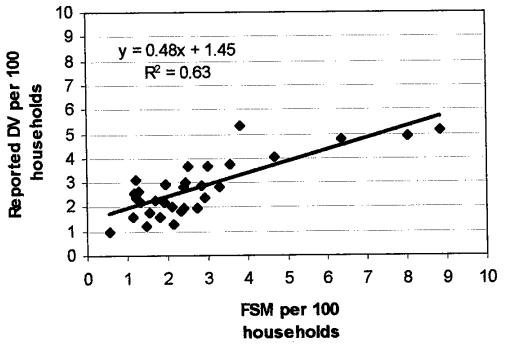


Figure 10: Regression of DV per 100 households on Income Support per100 households at ward level.

To strengthen any of these findings we would need to extend either the period covered by the data or cover a larger area (e.g. two local councils instead of one). Note also there may be other, more strongly associated factors, which we might have included and these may result in further refinements. For example, in separate analyses of reported DV incidents we found that certain ranges are more at risk than others. In particular we found the median age of victim was 30 years with an inter-quartile range of 25 to 42 years, whereas the median age of the accused was 36 (30 to 42 years). This is higher than typical ages reported in crime surveys and is probably a reflection of the fact we are dealing here with incidents reported to the police, which are known to have a higher

aggregations are meaningful, such as plots of households with 2 or more factors present. Alternatively the maps can be drawn to reflect patterns at a neighbourhood level using whatever geographical boundaries are deemed to be appropriate. The map in Figure 11, for example, shows neighbourhoods (600 m^2 *x* 600 m^2), which were chosen this size in order to be small enough to differentiate below ward level and yet large enough to contain sufficient observations.

Each neighbourhood is shaded according to whether they fall into the top decile of occurrence on each of four measures – FSM, DV, drug offending and mental health (noise complaints have been omitted since there are too few to create a meaningful map at neighbourhood level). The darker the cell the more factors apply. The results are remarkable since they show a clear concentration in the south of the borough corresponding to localities where DV is known to be highest (see Figure 7), but also some unexpected scattered pockets of problem neighbourhoods elsewhere. Interpretation is important however, since the analysis has shown DV can be prevalent where no risk factors are prominent. Nevertheless, a detailed comparison with Figure 7 indicates that Figure 11 is accurate as far as it goes, and certainly conveys more useful information than more standard ward-based maps.

4. Discussion and conclusions

It has been shown the approach discussed in this paper permits a detailed consideration of the vulnerability of different household types to various health and social risks and the locations in which they tend to concentrate. At the very minimum, it has also shown there is much information 'locked up' in administrative and publicly available data sets that could be exploited more than it is. The methodology itself is capable of development and work is continuing, for example, on modelling and related issues using regression techniques. More generally the methodology has also been shown to be useful for dealing with long-standing research problems, such as the ecological fallacy and the modifiable areal unit problem.

One productive line of enquiry has been analyzing the *ordering* of risk as well as the likelihood of particular factor combinations. An example currently under investigation is whether it is usual for school exclusions to be linked with teenage crime and if so whether criminal behavior has a tendency to occur before exclusion or after. The number of separately identifiable risk categories increases even more rapidly in these cases according to $\sum_{m=0}^{m=N} \frac{N!}{(N-m)!}$ so that whereas in the unordered cases five factors would produce 32 conditional risk measures (2^5), the ordered case would produce 326 measures. As a result systematic methods are needed to group the diluted risk pool into meaningful and statistically significant groupings.

It would be premature to say definitely whether the approach will unlock a greater understanding of interactions between the health and social economy and so lead to better predictive models, although the results from this case study seem promising. They show such phenomena are not distributed randomly but that there are systematic associations

between the risk factors considered. Whilst this is hardly a new finding the fact the methodology has enabled detailed quantification of different risk combinations, at the local as well as large scale, must be considered an improvement over comparable methods. At the same time it has opened the way to a more systematic evaluation of risk patterns and potential regularities or associations.

The case study described here is one of several carried to date and in all cases the results have found something new or unexpected to say about an issue or an area. However, more work is needed to obtain a greater understanding of regularities and patterns by for example comparing the same phenomena in different time windows or by expanding the population to cover adjacent councils. Whilst there is obvious scientific logic in taking forward such an agenda, it is important not to disregard the immediate potential applications for which the methodology can add value. These divide in three broad classes which we loosely call local area analysis, policy evaluation and, finally, resource allocation and cost analysis.

Local area analysis is in increasing demand for making cases to central Government or the European Commission for funding regeneration and other initiatives. As such initiatives become more targeted and focussed so the demand for detailed information and analysis increases. The NHE has already been used one such successful exercise and proved useful in changing perceptions about what the health needs were in a particular locality. In terms of policy analysis Governments are increasingly demanding an evidence base for policy interventions, but as far as local initiatives are concerned this has proved problematic for a range of reasons, most often data inadequacies and this is precisely where the NHE offers a solution. For example, imagine a scenario where systematic analyses take place at different points in time and across different neighbourhoods. The kind of approach described could help to ascertain whether a policy intervention has been effective or not.

The final class of applications concerns issues around resource allocation. Currently this can be very hit and miss at local geographical scales. Health workers do not always have a clear picture of local needs and so determining training needs or prioritizing services is an inexact science to a large extent. Similarly with so many agencies operating in an area it becomes virtually impossible to know what the services are costing overall, or the extent of any duplication of effort particularly in gathering and processing information that would allow greater co-ordination of resources. As part of the HAZ work some methods have already been tested at household level rather than using area based measures and so far results have been very encouraging.

The techniques, it seems, also have potential applications that go beyond the scope of this paper but a small footnote is worthy of mention. For example, if all data were assembled in this way dependence on centrally produced ward-based statistics might, in theory, be significantly reduced. Local authorities could construct their own data sets from their administrative data into whatever geographical units were deemed appropriate, if necessary under central guidance. It seems clear that statistical information obtained this way would also be more up to date than equivalent statistics currently disseminated by

central government, which are often years out of date. The question is to determine which Government statistics would be produced more quickly and reliably by this method, what new statistics could be published that are not currently available, and what the conventions would be about making them generally available for research purposes.

References

Acheson D (1998) Independent Inquiry into Inequalities in Health. The Stationery Office, UK.

Armitage, P and Berry, G (1987) Statistical Methods in Medical Research. Blackwell, Oxford.

Altman, D.G.(1999) Practical Statistics for Medical Research. Chapman & Hall/CRC, London..

Barnett, V (1984) Elements of Sampling Theory. Hodder and Stoughton, London.

Chou, Y (1972) Probability and Statistics for Decision Making. Holt, Reinhart and Winston, New York.

Freund J.E, (1973) Modern Elementary Statistics. Prentice-Hall International, London.

Greenland S, Robins J (1994) Invited commentary: ecological studies – biases, misconceptions, and counter examples. American Journal of Epidemiology, 139: 742-60.

Openshaw S, and Taylor P J (1981) 'The Modifiable Areal Unit Problem', in N. Wrigley and R.J. Bennett (editors), Quantitative Geography: A British View, Routledge, London.

Robinson W S (1950) Ecological correlations and the behavior of individuals. American Sociological Review, 15: 351-57.

Acknowledgements

The author is grateful for the continuing support of Brent Health Action Zone, and to Dr Martin Frost and Gillian Harper at Birkbeck College Department of Geography, and to the Brent GIS section in the Brent Council, particular Michelle Colley. He is also grateful to colleagues in the Department of Actuarial Sciences and Statistics at City University, London for their comments and advice, and to overseas colleagues such as Dr. Martin Spielauer in the Austrian Family Studies Institute, Vienna, and Dr Robert Gibberd of the University of Newcastle, NSW, Australia.

Annex A: Data preparation

Introduction – Land Parcel Identifier (LPI)

The basis of the methodology rests on the ability of being able to join together different data sets using address fields. In order to join data as accurately and efficiently as possible an address's unique land parcel identifier (LPI) is used. The property database, maintained by the London Borough of Brent contains a different record (including an LPI) for each property in Brent (a total of around 117,000 records).

Data Cleaning

Data sets used in this study were provided by various organizations and tended to be in different formats and of varying quality. Often within these organizations the practice of adhering to address standard BS7666 when collecting data has been slow to penetrate. As a result address information often required conversion to a standardized form, similar to that used in the property database, to enable each address's unique LPI to be retrieved and matched. This process is known as 'data cleaning' and involves importing the data into Microsoft Excel and using it to separate the complete postcode from the rest of the address string.

Address Matching Programme

An address-matching programme written in Visual Basic by the Department of Geography, Birkbeck College, is used to link data sets to the property database. This works by inputting a text file of the postcodes and address strings for each data-set. A typical data set could comprise a service delivered to that address or some other characterizing factor like free school meals. The programme checks the postcode first against the property database. If a match is found, it then goes on to check the next 17 characters of the address string. When an entire address is matched in this way, the address and its unique identifier from the property database is extracted to a 'matched' output file. Remaining unmatched addresses are then extracted to an 'unmatched' file. This process is found to have approximately a 70 – 80% success rate.

Those addresses unmatched to the property database using the address-matching programme have to be matched manually. In any given data set a few records tend to not be found at all, due mainly to some addresses being incorrectly inputted in the original data set.

Factor Combination Tables

By now each service data set will comprise just a list of LPIs. There is no need to have included any of the other address information as the LPI has taken the place of the 'text address'. Similarly all other information from the original data set has been 'stripped-off'. In this way the data is anonymized, so that LPI combinations can be analyzed without referring to individuals or their original addresses.

Each data set and an extract from the property database of all LPIs within the Borough are stored as separate tables in Access. Any duplicate LPIs within individual service data sets are removed using a 'remove duplicates' query and the number "1" inputted in a separate column alongside each remaining LPIs. From these separate data sets it is possible to create a binary matrix of received services.

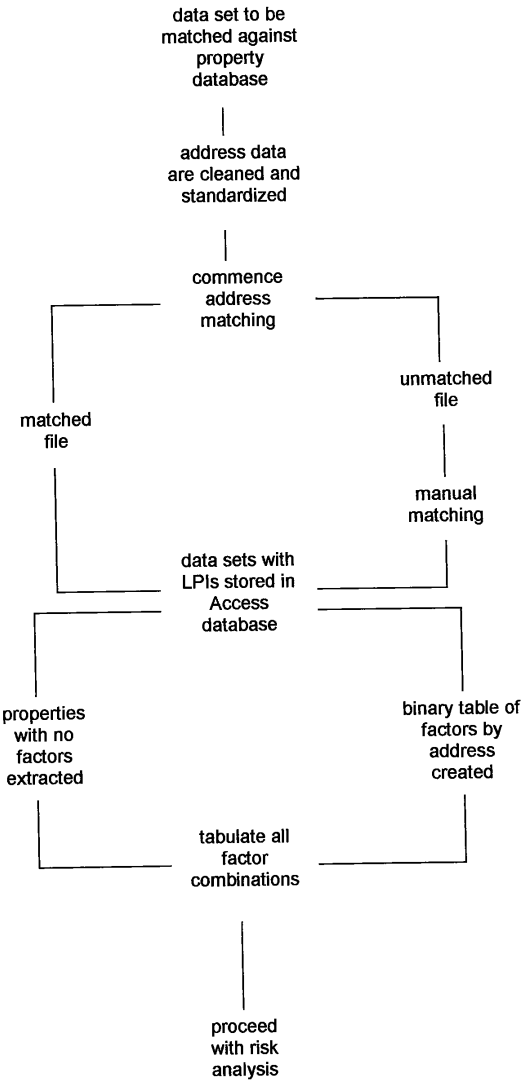
The property database extract, a list of every LPI in the borough, forms the basis of the table. The various service data sets are added to this 'spine' using a query in Access creating a binary matrix, as the example below illustrates:

LPI	Factor 1	factor 2	factor 3	factor 4	factor...n
LPI 1	1	0	0	0	0
LPI 2	1	0	1	1	0
LPI 3	0	0	0	0	0
LPI 4	0	0	0	1	0
LPI 5	0	0	0	0	0
LPI 6	1	1	1	1	1
LPI...m	1	1	0	1	0

Those properties (LPIs) not attracting any factor are extracted from the list (LPIs 3 & 5 in the example above). These records are addresses can be referred to as the 'no factor' matrix because no service or factor has been matched to their addresses. Removing these records reduces the size of the list and speeds the service overlap analysis. All the remaining LPIs will have a "1" mark in at least one of the service columns, this can be referred to as the 'factor matrix'.

The next stage is to identify the combination of services received within each property address (LPI). Again this done in Access, creating a 'combinations matrix' which lists each possible combination of factors from which the results are transferred to a summary sheet. This sheet lists all occurrences of each service and each possible combination. The occurrences of combinations are then analyzed in relation to calculated probabilities.

The whole process is illustrated in the accompanying diagram.


data set to be matched against property database
address data are cleaned and standardized
commence address matching
matched file
unmatched file
manual matching
data sets with LPIs stored in Access database
properties with no factors extracted
binary table of factors by address created
tabulate all factor combinations
proceed with risk analysis

Annex B: DV risk ladders and ready reckoners– observed versus predicted using Ordinary Least Squares (OLS) and Maximum Likelihood (ML)

level	ABCDE	No. of households	observed risk of DV %	predicted risk of DV (OLS)	predicted risk of DV (ML)
3	10101	2	100.0	13.7	25.5
2	01010	3	33.3	17.4	12.6
2	00011	19	26.3	21.8	17.0
3	00111	34	20.6	23.3	29.7
3	01101	15	13.3	13.5	20.6
2	00110	323	11.8	10.4	12.0
3	01110	10	10.0	18.7	23.0
2	00101	1265	9.6	7.3	10.6
2	01001	11	9.1	12.5	11.2
1	00010	405	6.9	9.6	6.2
2	01100	366	6.8	5.6	7.7
1	00001	1209	6.7	6.8	5.5
2	10100	35	5.7	5.7	10.0
1	01000	374	5.3	5.2	3.9
1	00100	23944	3.8	2.9	3.7
0	00000	74356	1.8	2.7	1.8

Table B.1: Any DV incident observed and predicted risk using a) OLS; b) maximum likelihood: Key: A) noise complaints; B) mental health; C) social housing; D) drug offences; E) free school meals

Level	ABCDE	No. of households	observed risk of DV %	predicted risk of DV (OLS)	predicted risk of DV (ML)
3	10101	2	100.0	15.1	28.1
2	01010	3	33.3	14.6	7.0
3	00111	34	11.8	12.2	16.3
2	00011	19	10.5	11.8	8.1
3	01110	10	10.0	15.1	14.1
3	01101	15	6.7	9.3	12.9
2	10100	35	5.7	5.7	9.8
2	00110	323	5.3	4.5	5.1
2	00101	1265	4.1	2.7	4.7
2	01100	366	3.0	3.4	4.0
1	01000	374	2.9	3.3	1.8
1	00001	1209	2.8	2.6	2.2
1	00010	405	2.5	4.4	2.4
1	00100	23944	1.4	0.9	1.3
0	00000	74356	0.6	0.9	0.6

Table B.2: Multiple DV incidents at each address observed and predicted using a) OLS; b) maximum likelihood

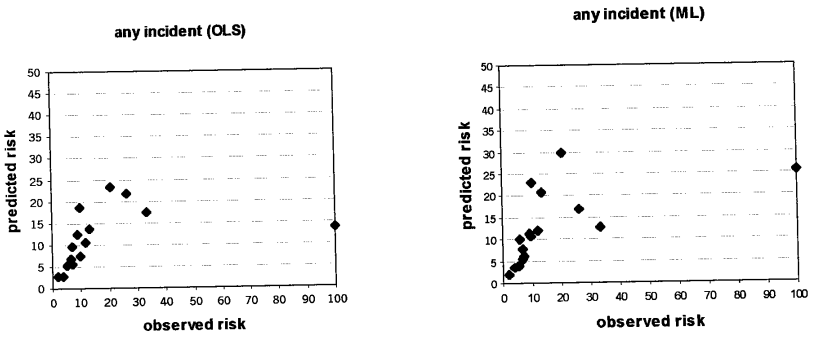


Figure B.1: Any incident observed versus predicted risk – OLS and ML

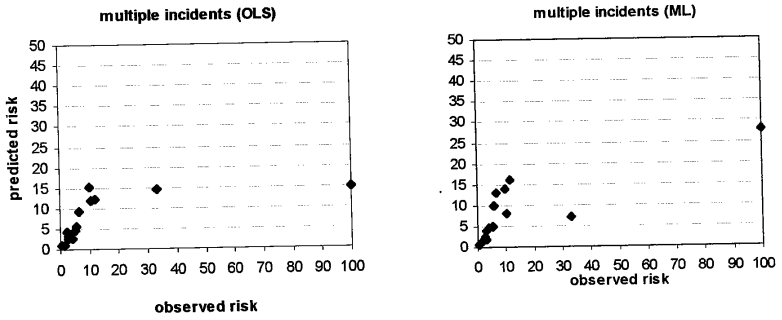


Figure B.2: Multiple incidents at same address observed versus predicted – OLS and ML

risk category	10101	01010	00011	00111	01101	00110	01110	00101	01001	00010	01100	00001	10100	01000
10101	*1.00*	3.00	3.80	4.86	7.50	8.50	10.00	10.37	11.00	14.46	14.64	14.93	17.50	18.70
01010	0.33	*1.00*	1.27	1.62	2.50	2.83	3.33	3.46	3.67	4.82	4.88	4.98	5.83	6.23
00011	0.26	0.79	*1.00*	1.28	1.97	2.24	2.63	2.73	2.89	3.81	3.85	3.93	4.61	4.92
00111	0.21	0.62	0.78	*1.00*	1.54	1.75	2.06	2.13	2.26	2.98	3.01	3.07	3.60	3.85
01101	0.13	0.40	0.51	0.65	*1.00*	1.13	1.33	1.38	1.47	1.93	1.95	1.99	2.33	2.49
00110	0.12	0.35	0.45	0.57	0.88	*1.00*	1.18	1.22	1.29	1.70	1.72	1.76	2.06	2.20
01110	0.10	0.30	0.38	0.49	0.75	0.85	*1.00*	1.04	1.10	1.45	1.46	1.49	1.75	1.87
00101	0.10	0.29	0.37	0.47	0.72	0.82	0.96	*1.00*	1.06	1.39	1.41	1.44	1.69	1.80
01001	0.09	0.27	0.35	0.44	0.68	0.77	0.91	0.94	*1.00*	1.31	1.33	1.36	1.59	1.70
00010	0.07	0.21	0.26	0.34	0.52	0.59	0.69	0.72	0.76	*1.00*	1.01	1.03	1.21	1.29
01100	0.07	0.20	0.26	0.33	0.51	0.58	0.68	0.71	0.75	0.99	*1.00*	1.02	1.20	1.28
00001	0.07	0.20	0.25	0.33	0.50	0.57	0.67	0.69	0.74	0.97	0.98	*1.00*	1.17	1.25
10100	0.06	0.17	0.22	0.28	0.43	0.49	0.57	0.59	0.63	0.83	0.84	0.85	*1.00*	1.07
01000	0.05	0.16	0.20	0.26	0.40	0.45	0.53	0.55	0.59	0.77	0.78	0.80	0.94	*1.00*

Table B.3: Ready reckoner of observed relative risk based on any DV incident

risk category	10101	01010	00011	00111	01101	00110	01110	00101	01001	00010	01100	00001	10100	01000
10101	1.00	3.00	8.50	9.50	10.00	15.00	17.50	19.00	24.33	33.27	34.00	35.56	40.50	71.90
01010	0.33	1.00	2.83	3.17	3.33	5.00	5.83	6.33	8.11	11.09	11.33	11.85	13.50	23.97
00011	0.12	0.35	1.00	1.12	1.18	1.76	2.06	2.24	2.86	3.91	4.00	4.18	4.76	8.46
00111	0.11	0.32	0.89	1.00	1.05	1.58	1.84	2.00	2.56	3.50	3.58	3.74	4.26	7.57
01101	0.10	0.30	0.85	0.95	1.00	1.50	1.75	1.90	2.43	3.33	3.40	3.56	4.05	7.19
00110	0.07	0.20	0.57	0.63	0.67	1.00	1.17	1.27	1.62	2.22	2.27	2.37	2.70	4.79
01110	0.06	0.17	0.49	0.54	0.57	0.86	1.00	1.09	1.39	1.90	1.94	2.03	2.31	4.11
00101	0.05	0.16	0.45	0.50	0.53	0.79	0.92	1.00	1.28	1.75	1.79	1.87	2.13	3.78
01001	0.04	0.12	0.35	0.39	0.41	0.62	0.72	0.78	1.00	1.37	1.40	1.46	1.66	2.96
00010	0.03	0.09	0.26	0.29	0.30	0.45	0.53	0.57	0.73	1.00	1.02	1.07	1.22	2.16
01100	0.03	0.09	0.25	0.28	0.29	0.44	0.51	0.56	0.72	0.98	1.00	1.05	1.19	2.11
00001	0.03	0.08	0.24	0.27	0.28	0.42	0.49	0.53	0.68	0.94	0.96	1.00	1.14	2.02
10100	0.02	0.07	0.21	0.23	0.25	0.37	0.43	0.47	0.60	0.82	0.84	0.88	1.00	1.78
01000	0.01	0.04	0.12	0.13	0.14	0.21	0.24	0.26	0.34	0.46	0.47	0.49	0.56	1.00

Table B.4: Ready reckoner of observed relative risk based on multiple DV incidents. For key to codes see caption Table B.1.

risk category	11111	11011	10111	01111	10011	01011	11110	00111	11101	11010	00011	11001	10110	10101	01110
11111	1.00	1.12	1.27	1.43	1.45	1.65	1.82	1.91	1.97	2.11	2.23	2.30	2.48	2.71	2.88
11011	0.89	1.00	1.13	1.28	1.29	1.47	1.62	1.71	1.76	1.89	1.99	2.05	2.22	2.42	2.57
10111	0.79	0.88	1.00	1.13	1.14	1.30	1.43	1.50	1.55	1.66	1.75	1.81	1.95	2.13	2.27
01111	0.70	0.78	0.89	1.00	1.01	1.15	1.27	1.33	1.38	1.48	1.56	1.61	1.73	1.89	2.01
10011	0.69	0.77	0.88	0.99	1.00	1.14	1.25	1.32	1.36	1.46	1.54	1.59	1.71	1.87	1.99
01011	0.61	0.68	0.77	0.87	0.88	1.00	1.10	1.16	1.20	1.28	1.35	1.40	1.51	1.65	1.75
11110	0.55	0.62	0.70	0.79	0.80	0.91	1.00	1.05	1.08	1.16	1.23	1.27	1.37	1.49	1.59
00111	0.52	0.59	0.66	0.75	0.76	0.86	0.95	1.00	1.03	1.11	1.17	1.20	1.30	1.42	1.51
11101	0.51	0.57	0.64	0.73	0.74	0.83	0.92	0.97	1.00	1.07	1.13	1.17	1.26	1.38	1.46
11010	0.47	0.53	0.60	0.68	0.69	0.78	0.86	0.90	0.93	1.00	1.05	1.09	1.17	1.28	1.36
00011	0.45	0.50	0.57	0.64	0.65	0.74	0.82	0.86	0.88	0.95	1.00	1.03	1.11	1.22	1.29
11001	0.43	0.49	0.55	0.62	0.63	0.72	0.79	0.83	0.86	0.92	0.97	1.00	1.08	1.18	1.25
10110	0.40	0.45	0.51	0.58	0.58	0.66	0.73	0.77	0.79	0.85	0.90	0.93	1.00	1.09	1.16
10101	0.37	0.41	0.47	0.53	0.53	0.61	0.67	0.70	0.73	0.78	0.82	0.85	0.92	1.00	1.06
01110	0.35	0.39	0.44	0.50	0.50	0.57	0.63	0.66	0.68	0.73	0.77	0.80	0.86	0.94	1.00
10010	0.34	0.38	0.43	0.49	0.50	0.56	0.62	0.65	0.67	0.72	0.76	0.79	0.85	0.93	0.99
01101	0.32	0.36	0.40	0.45	0.46	0.52	0.58	0.61	0.62	0.67	0.71	0.73	0.79	0.86	0.91
10001	0.31	0.35	0.40	0.45	0.45	0.51	0.57	0.60	0.62	0.66	0.70	0.72	0.78	0.85	0.90
01010	0.29	0.33	0.37	0.42	0.43	0.48	0.53	0.56	0.58	0.62	0.65	0.68	0.73	0.80	0.85
01001	0.27	0.30	0.34	0.38	0.39	0.44	0.49	0.51	0.53	0.57	0.60	0.62	0.68	0.73	0.77
00110	0.25	0.28	0.31	0.35	0.36	0.41	0.45	0.47	0.49	0.52	0.55	0.57	0.61	0.67	0.71
11100	0.24	0.27	0.30	0.34	0.34	0.39	0.43	0.45	0.47	0.50	0.53	0.55	0.59	0.64	0.68
00101	0.22	0.25	0.28	0.32	0.32	0.37	0.41	0.43	0.44	0.47	0.50	0.52	0.56	0.61	0.64
00010	0.21	0.23	0.26	0.30	0.30	0.34	0.38	0.39	0.41	0.44	0.46	0.47	0.51	0.56	0.59
11000	0.20	0.22	0.25	0.29	0.29	0.33	0.36	0.38	0.39	0.42	0.44	0.46	0.49	0.54	0.57
00001	0.19	0.21	0.24	0.27	0.27	0.31	0.34	0.36	0.37	0.40	0.42	0.43	0.47	0.51	0.54
10100	0.17	0.19	0.21	0.24	0.24	0.27	0.30	0.32	0.33	0.35	0.37	0.38	0.41	0.45	0.48
01100	0.14	0.16	0.18	0.20	0.20	0.23	0.25	0.26	0.27	0.29	0.31	0.32	0.35	0.38	0.40
10000	0.14	0.15	0.18	0.20	0.20	0.23	0.25	0.26	0.27	0.29	0.31	0.32	0.34	0.37	0.40
01000	0.12	0.13	0.15	0.17	0.17	0.19	0.21	0.22	0.23	0.25	0.26	0.27	0.29	0.32	0.34
00100	0.10	0.11	0.12	0.14	0.14	0.16	0.18	0.18	0.19	0.20	0.22	0.22	0.24	0.26	0.28
00000	0.08	0.09	0.10	0.11	0.12	0.13	0.15	0.15	0.16	0.17	0.18	0.18	0.20	0.22	0.23

risk category	01101	10001	01010	01001	00110	11100	00101	00010	11000	00001	10100	01100	10000	01000	00100
11111	3.15	3.20	3.41	3.74	4.06	4.21	4.47	4.84	5.02	5.34	6.03	7.12	7.24	8.57	10.36
11011	2.81	2.86	3.04	3.34	3.63	3.76	3.99	4.32	4.48	4.76	5.39	6.36	6.46	7.65	9.25
10111	2.48	2.52	2.68	2.94	3.20	3.31	3.52	3.81	3.95	4.20	4.75	5.61	5.70	6.74	8.16
01111	2.20	2.24	2.38	2.61	2.84	2.94	3.12	3.38	3.51	3.73	4.21	4.98	5.06	5.98	7.24
10011	2.18	2.21	2.35	2.58	2.80	2.90	3.08	3.34	3.47	3.68	4.16	4.92	5.00	5.91	7.15
01011	1.92	1.95	2.07	2.27	2.47	2.56	2.71	2.94	3.05	3.24	3.67	4.33	4.40	5.21	6.30
11110	1.74	1.76	1.88	2.06	2.24	2.32	2.46	2.67	2.76	2.94	3.32	3.92	3.99	4.71	5.70
00111	1.65	1.67	1.78	1.96	2.13	2.20	2.34	2.53	2.63	2.79	3.16	3.73	3.79	4.48	5.42
11101	1.60	1.62	1.73	1.90	2.06	2.14	2.27	2.46	2.55	2.71	3.06	3.61	3.67	4.35	5.26
11010	1.49	1.51	1.61	1.77	1.92	1.99	2.11	2.29	2.38	2.52	2.85	3.37	3.43	4.05	4.90
00011	1.42	1.44	1.53	1.68	1.82	1.89	2.00	2.17	2.25	2.39	2.71	3.20	3.25	3.84	4.65
11001	1.37	1.39	1.48	1.63	1.77	1.83	1.94	2.11	2.18	2.32	2.62	3.10	3.15	3.72	4.50
10110	1.27	1.29	1.37	1.51	1.64	1.70	1.80	1.95	2.02	2.15	2.43	2.87	2.92	3.45	4.17
10101	1.16	1.18	1.26	1.38	1.50	1.55	1.65	1.79	1.85	1.97	2.23	2.63	2.67	3.16	3.82
01110	1.09	1.11	1.18	1.30	1.41	1.46	1.55	1.68	1.74	1.85	2.09	2.47	2.51	2.97	3.60
10010	1.08	1.09	1.17	1.28	1.39	1.44	1.53	1.66	1.72	1.83	2.06	2.44	2.48	2.93	3.54
01101	1.00	1.02	1.08	1.19	1.29	1.33	1.42	1.54	1.59	1.69	1.91	2.26	2.30	2.72	3.29
10001	0.99	1.00	1.06	1.17	1.27	1.31	1.40	1.51	1.57	1.67	1.88	2.23	2.26	2.68	3.24
01010	0.93	0.94	1.00	1.10	1.19	1.24	1.31	1.42	1.47	1.57	1.77	2.09	2.13	2.51	3.04
01001	0.84	0.86	0.91	1.00	1.09	1.13	1.19	1.30	1.34	1.43	1.61	1.91	1.94	2.29	2.77
00110	0.78	0.79	0.84	0.92	1.00	1.04	1.10	1.19	1.24	1.31	1.49	1.75	1.78	2.11	2.55
11100	0.75	0.76	0.81	0.89	0.97	1.00	1.06	1.15	1.19	1.27	1.43	1.69	1.72	2.04	2.46
00101	0.71	0.72	0.76	0.84	0.91	0.94	1.00	1.08	1.12	1.19	1.35	1.59	1.62	1.92	2.32
00010	0.65	0.66	0.70	0.77	0.84	0.87	0.92	1.00	1.04	1.10	1.25	1.47	1.49	1.77	2.14
11000	0.63	0.64	0.68	0.74	0.81	0.84	0.89	0.96	1.00	1.06	1.20	1.42	1.44	1.71	2.06
00001	0.59	0.60	0.64	0.70	0.76	0.79	0.84	0.91	0.94	1.00	1.13	1.34	1.36	1.61	1.94
10100	0.52	0.53	0.56	0.62	0.67	0.70	0.74	0.80	0.83	0.88	1.00	1.18	1.20	1.42	1.72
01100	0.44	0.45	0.48	0.52	0.57	0.59	0.63	0.68	0.71	0.75	0.85	1.00	1.02	1.20	1.45
10000	0.44	0.44	0.47	0.52	0.56	0.58	0.62	0.67	0.69	0.74	0.83	0.98	1.00	1.18	1.43
01000	0.37	0.37	0.40	0.44	0.47	0.49	0.52	0.57	0.59	0.62	0.70	0.83	0.85	1.00	1.21
00100	0.30	0.31	0.33	0.36	0.39	0.41	0.43	0.47	0.48	0.52	0.58	0.69	0.70	0.83	1.00
00000	0.25	0.26	0.27	0.30	0.32	0.34	0.36	0.39	0.40	0.43	0.48	0.57	0.58	0.69	0.83

Table B.5: Ready reckoner of predicted relative risk based on any DV incident using OLS. For key to codes see Table B.1.

Cass Business School
City of London

FACULTY OF ACTUARIAL SCIENCE AND STATISTICS

Actuarial Research Papers since 2001

---

135. Renshaw A. E. and Haberman S. On the Forecasting of Mortality Reduction Factors. February 2001.
ISBN 1 901615 56 1

136. Haberman S., Butt Z. & Rickayzen B. D. Multiple State Models, Simulation and Insurer Insolvency. February 2001. 27 pages.
ISBN 1 901615 57 X

137. Khorasanee M.Z. A Cash-Flow Approach to Pension Funding. September 2001. 34 pages.
ISBN 1 901615 58 8

138. England P.D. Addendum to "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". November 2001. 17 pages.
ISBN 1 901615 59 6

139. Verrall R.J. A Bayesian Generalised Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. November 2001. 10 pages.
ISBN 1 901615 62 6

140. Renshaw A.E. and Haberman. S. Lee-Carter Mortality Forecasting, a Parallel GLM Approach, England and Wales Mortality Projections. January 2002. 38 pages.
ISBN 1 901615 63 4

141. Ballotta L. and Haberman S. Valuation of Guaranteed Annuity Conversion Options. January 2002. 25 pages.
ISBN 1 901615 64 2

142. Butt Z. and Haberman S. Application of Frailty-Based Mortality Models to Insurance Data. April 2002. 65 pages.
ISBN 1 901615 65 0

143. Gerrard R.J. and Glass C.A. Optimal Premium Pricing in Motor Insurance: A Discrete Approximation. (Will be available 2003).

144. Mayhew, L. The Neighbourhood Health Economy. A systematic approach to the examination of health and social risks at neighbourhood level. December 2002. 43 pages.
ISBN 1 901615 66 9

Statistical Research Papers

1. Sebastiani P. Some Results on the Derivatives of Matrix Functions. December 1995. 17 Pages.
ISBN 1 874 770 83 2

2. Dawid A.P. and Sebastiani P. Coherent Criteria for Optimal Experimental Design. March 1996. 35 Pages.
ISBN 1 874 770 86 7

3. Sebastiani P. and Wynn H.P. Maximum Entropy Sampling and Optimal Bayesian Experimental Design. March 1996. 22 Pages.
ISBN 1 874 770 87 5

4. Sebastiani P. and Settimi R. A Note on D-optimal Designs for a Logistic Regression Model. May 1996. 12 Pages.
ISBN 1 874 770 92 1

5. Sebastiani P. and Settimi R. First-order Optimal Designs for Non Linear Models. August 1996. 28 Pages.
ISBN 1 874 770 95 6

6. Newby M. A Business Process Approach to Maintenance: Measurement, Decision and Control. September 1996. 12 Pages.
ISBN 1 874 770 96 4

7. Newby M. Moments and Generating Functions for the Absorption Distribution and its Negative Binomial Analogue. September 1996. 16 Pages.
ISBN 1 874 770 97 2

8. Cowell R.G. Mixture Reduction via Predictive Scores. November 1996. 17 Pages.
ISBN 1 874 770 98 0

9. Sebastiani P. and Ramoni M. Robust Parameter Learning in Bayesian Networks with Missing Data. March 1997. 9 Pages.
ISBN 1 901615 00 6

10. Newby M.J. and Coolen F.P.A. Guidelines for Corrective Replacement Based on Low Stochastic Structure Assumptions. March 1997. 9 Pages.
ISBN 1 901615 01 4.

11. Newby M.J. Approximations for the Absorption Distribution and its Negative Binomial Analogue. March 1997. 6 Pages.
ISBN 1 901615 02 2

12. Ramoni M. and Sebastiani P. The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases. June 1997. 11 Pages.
ISBN 1 901615 10 3

13. Ramoni M. and Sebastiani P. Learning Bayesian Networks from Incomplete Databases. June 1997. 14 Pages.
ISBN 1 901615 11 1

14. Sebastiani P. and Wynn H.P. Risk Based Optimal Designs. June 1997. 10 Pages.
ISBN 1 901615 13 8

15. Cowell R. Sampling without Replacement in Junction Trees. June 1997. 10 Pages.
ISBN 1 901615 14 6

16. Dagg R.A. and Newby M.J. Optimal Overhaul Intervals with Imperfect Inspection and Repair. July 1997. 11 Pages. ISBN 1 901615 15 4

17. Sebastiani P. and Wynn H.P. Bayesian Experimental Design and Shannon Information. October 1997. 11 Pages. ISBN 1 901615 17 0

18. Wolstenholme L.C. A Characterisation of Phase Type Distributions. November 1997. 11 Pages. ISBN 1 901615 18 9

19. Wolstenholme L.C. A Comparison of Models for Probability of Detection (POD) Curves. December 1997. 23 Pages. ISBN 1 901615 21 9

20. Cowell R.G. Parameter Learning from Incomplete Data Using Maximum Entropy I: Principles. February 1999. 19 Pages. ISBN 1 901615 37 5

21. Cowell R.G. Parameter Learning from Incomplete Data Using Maximum Entropy II: Application to Bayesian Networks. November 1999. 12 Pages ISBN 1 901615 40 5

22. Cowell R.G. FINEX : Forensic Identification by Network Expert Systems. March 2001. 10 pages. ISBN 1 901615 60X

23. Cowell R.G. When Learning Bayesian Networks from Data, using Conditional Independence Tests is Equivalant to a Scoring Metric. March 2001. 11 pages. ISBN 1 901615 61 8

Faculty of Actuarial Science and Statistics

Actuarial Research Club

The support of the corporate members

CGNU Assurance
Computer Sciences Corporation
English Matthews Brockman
Government Actuary's Department
HCM Consultants (UK) Ltd
KPMG
PricewaterhouseCoopers
Swiss Reinsurance
Watson Wyatt Partners

is gratefully acknowledged.

ISBN 1 901615 66 9